The new CMS DAQ system for LHC operation after 2014 (DAQ2)

CHEP2013: Computing in High Energy Physics 2013
14-18 Oct 2013
Amsterdam

Andre Holzner, University California at San Diego

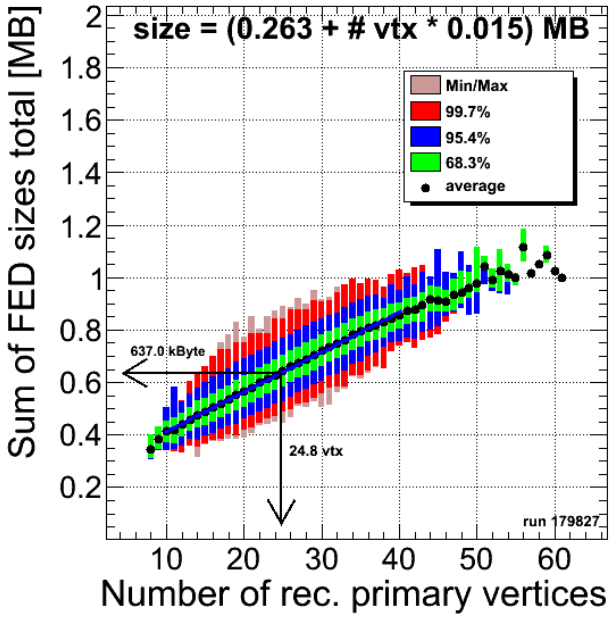
On behalf of the CMS collaboration

Overview

- DAQ2 Motivation
- Requirements
- Layout / Data path
- Frontend Readout Link
- Event builder core
- Performance considerations
- Infiniband
- File based filter farm and storage
- DAQ2 test setup and results
- Summary/Outlook

DAQ2 motivation

- Aging equipment:
  - Run1 DAQ uses some technologies which are disappearing
    - PCI-X cards, Myrinet
    - Almost all equipment reached the end of the 5 year lifecycle
- CMS detector upgrades
  - Some subsystems move to new front-end drivers
  - Some subsystems will add more channels
- LHC performance
  - Expect higher instantaneous luminosity after LS1
    → higher number of interactions per bunch crossing ('pileup')
    → larger event size, higher data rate
- Physics
  - Higher centre-of-mass energy and more pileup imply:
    - either raise trigger thresholds, or
    - more intelligent decisions at Higher Level Trigger
      → requires more CPU power

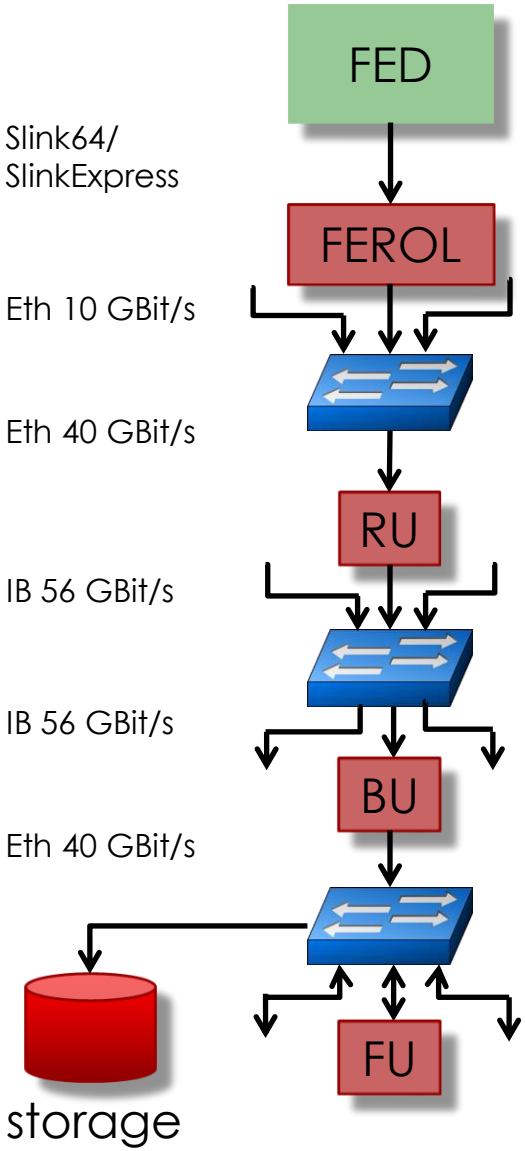


DAQ2 requirements

Requirement	DAQ1	DAQ2
Readout rate	100 kHz	
Front end drivers (FEDs)	640: 1 ~ 2 kByte	640: 1 ~ 2 kByte ~50: 2 ~ 8 kByte
Total readout bandwidth	100 GByte/s	200 GByte/s
Interface to FEDs[1)]	SLink64	Slink64/Slink Express
Coupling Event builder software/HLT software[2)]	no requirement	decoupled
Lossless event building	□	
HLT capacity	extendable	
High availabilty/ fault tolerance[3)]	□	
Cloud facility for offline processing[4)]	originally not required	□
Subdetector local runs	□	

See talks of 1) P. Žejdl 2) R. Mommsen 3) H.Sakulin 4) J.A.Coarasa

DAQ2 data path



~640 (legacy) + 50 (µTCA) Front End Drivers

~576 Front End Readout Optical Links (FEROLs)

custom hardware

10 GBit/s Ethernet → 40 GBit/s Ethernet switches, 8/12/16 →1 concentration

72 Readout Unit PCs (superfragment assembly)

Infiniband switch (full 72 x 48 connectivity, 2.7 TBit/s)

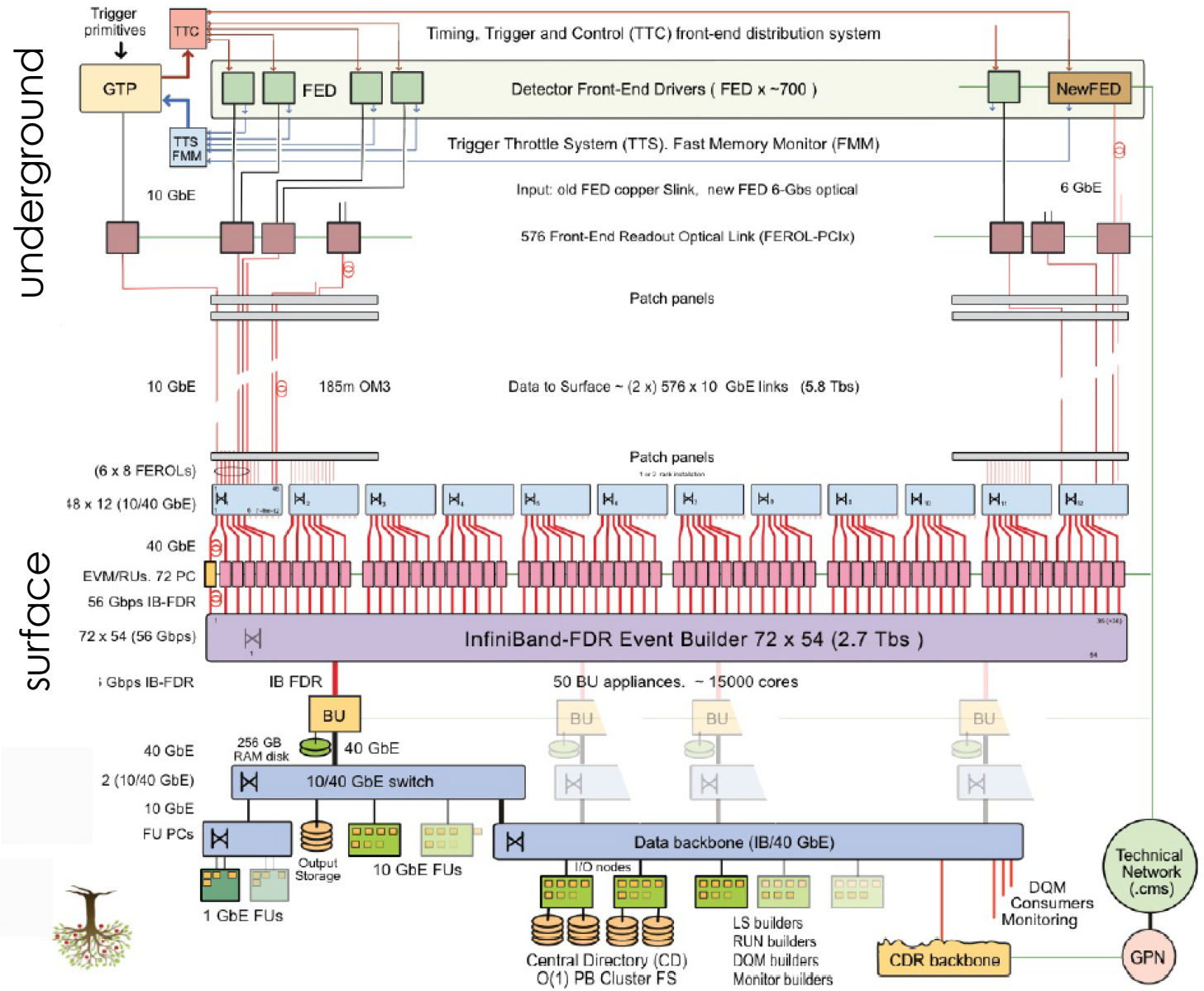
48 Builder units (full event assembly)

Ethernet switches 40 GBit/s → 10 GBit/s (→ 1 GBit/s), 1 →M distribution

Filter units (~ 13'000 cores)

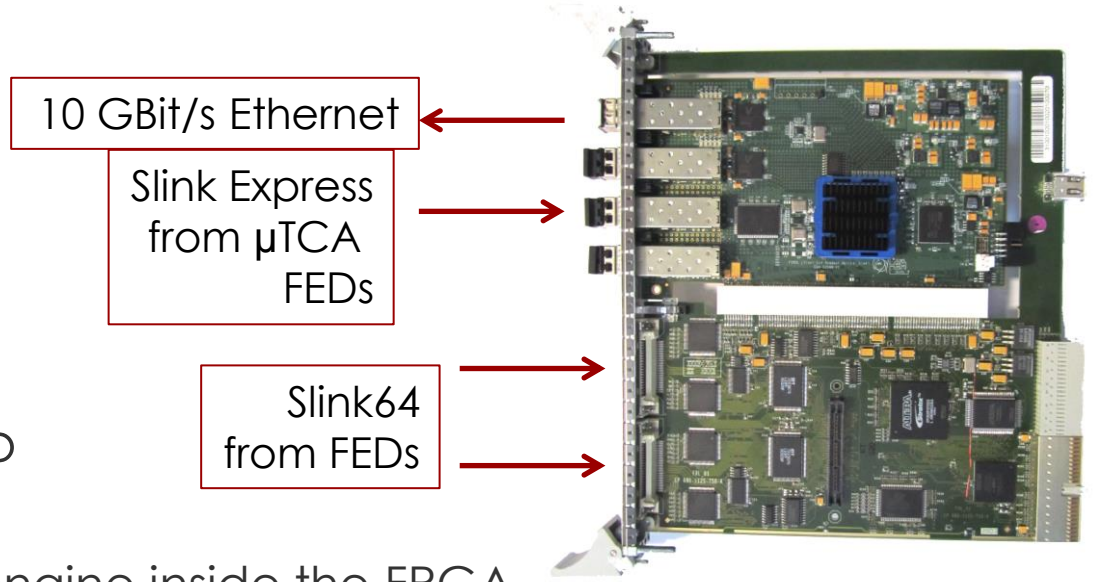
commercial hardware

DAQ2 layout


underground
surface
Trigger primitives
TTC
GTP
TTS FMM
FED
NewFED
Timing, Trigger and Control (TTC) front-end distribution system
Detector Front-End Drivers (FED x ~700)
Trigger Throttle System (TTS). Fast Memory Monitor (FMM)
Input: old FED copper Slink, new FED 6-Gbs optical
576 Front-End Readout Optical Link (FEROL-PCIx)
10 GbE
6 GbE
Patch panels
10 GbE
185m OM3
Data to Surface ~ (2 x) 576 x 10 GbE links (5.8 Tbs)
Patch panels
(6 x 8 FEROLs)
8 x 12 (10/40 GbE)
40 GbE
EVM/RUs. 72 PC
56 Gbps IB-FDR
72 x 54 (56 Gbps)
InfiniBand-FDR Event Builder 72 x 54 (2.7 Tbs)
Gbps IB-FDR
IB FDR
50 BU appliances. ~ 15000 cores
BU
40 GbE
256 GB RAM disk
40 GbE
2 (10/40 GbE)
10/40 GbE switch
10 GbE
FU PCs
Data backbone (IB/40 GbE)
Output Storage
10 GbE FUs
1 GbE FUs
I/O nodes
Central Directory (CD)
O(1) PB Cluster FS
LS builders
RUN builders
DQM builders
Monitor builders
CDR backbone
DQM Consumers Monitoring
Technical Network (.cms)
GPN

FrontEnd Readout Link (FEROL)

- Replace Myrinet card (upper half) by a new custom card
- PCI-X interface to legacy slink receiver card (lower half)
- 10 GBit/s Ethernet output to central event builder
  - Restricted TCP/IP protocol engine inside the FPGA



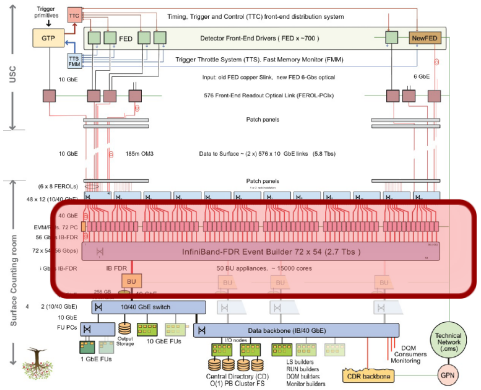
- Additional optical links (inputs) for future µTCA based Front End Drivers (6-10 GBit/s; custom, simple point to point protocol)

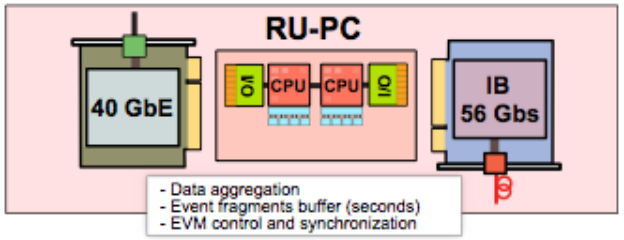
- Allows to use industry standard 10 GBit/s transceivers, cables and switches/routers
  - Only commercially available hardware further downstream

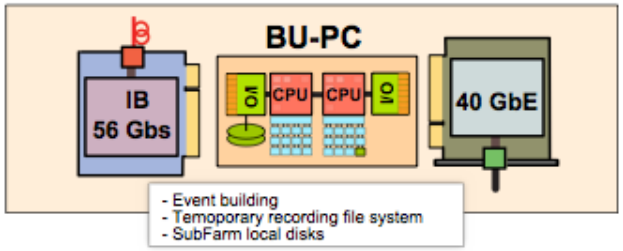
see P. Žejdl's talk for more details

Event Builder Core

- Two stage event building:
  - 72 Readout units (RU) aggregate 8-16 fragments (4 kByte average) into superfragments
    - Larger buffers compared to FEROLs
  - 48 Builder Units (BU) build the entire event from superfragments
- InifiniBand (or 40 Gbit/s Ethernet) as interconnect
- Works in a 15 x 15 system, need to scale to 72 x 48
- Fault tolerance:
  - FEROLs can be routed to different RU (adding a second switching layer improves flexibility)
  - Builder Units can be excluded from running





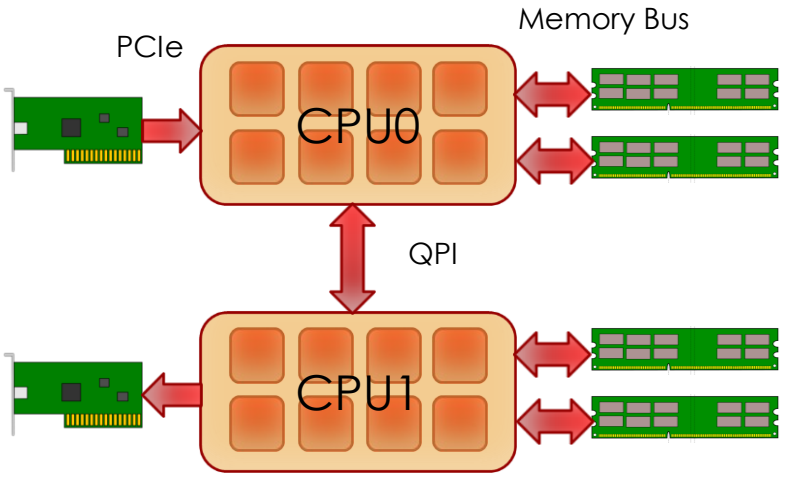


Performance considerations

- Number of DAQ2 elements is an order of magnitude smaller than for DAQ1
- Consequently, bandwidth per PC is an order of magnitude higher

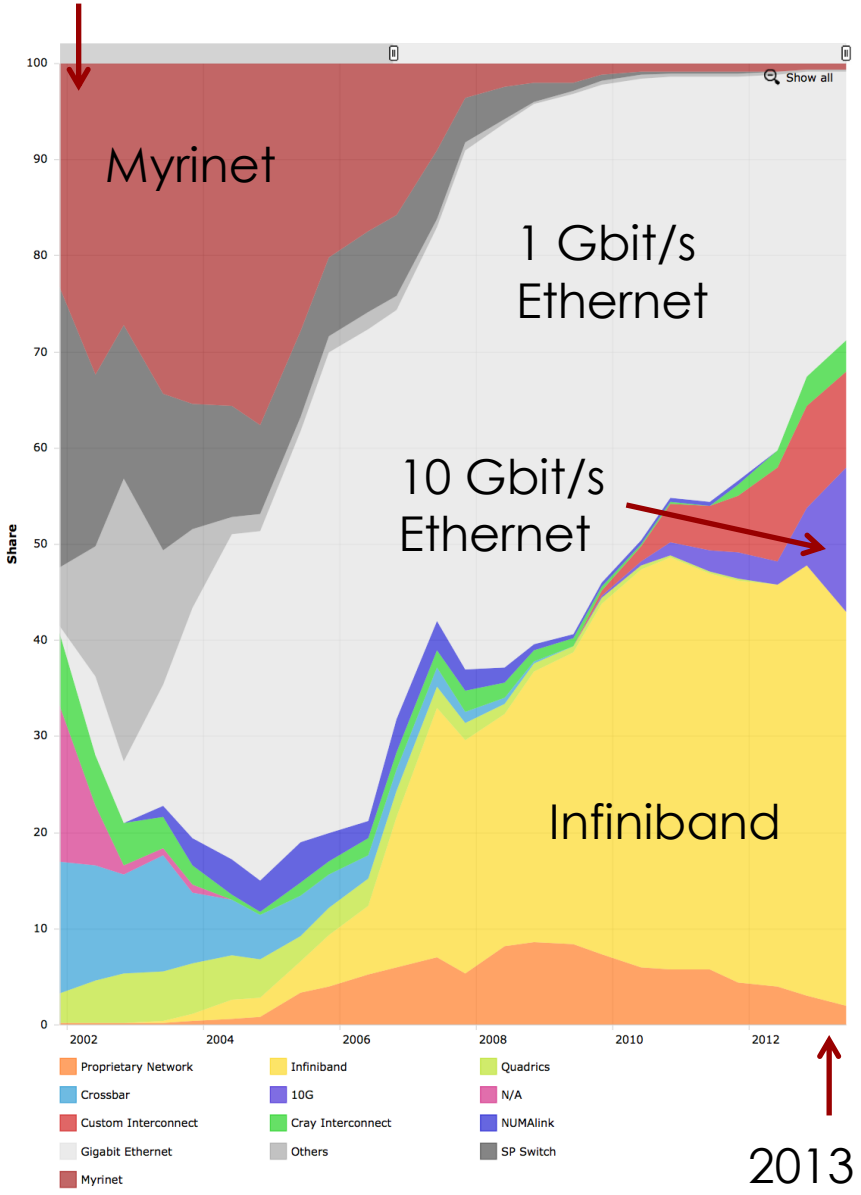
	DAQ1	DAQ2
# readout units (RU)	640	48
RU max. bandwidth	3 Gbit/s	40 Gbit/s
# builder units (BU)	>1000	72
BU max. bandwidth	2 Gbit/s	56 Gbit/s

- CPU frequency did not increase since DAQ1 but number of cores did
- Need to pay attention to performance tuning
  - TCP socket buffers
  - Interrupt affinities
  - Non-uniform memory access



Infiniband

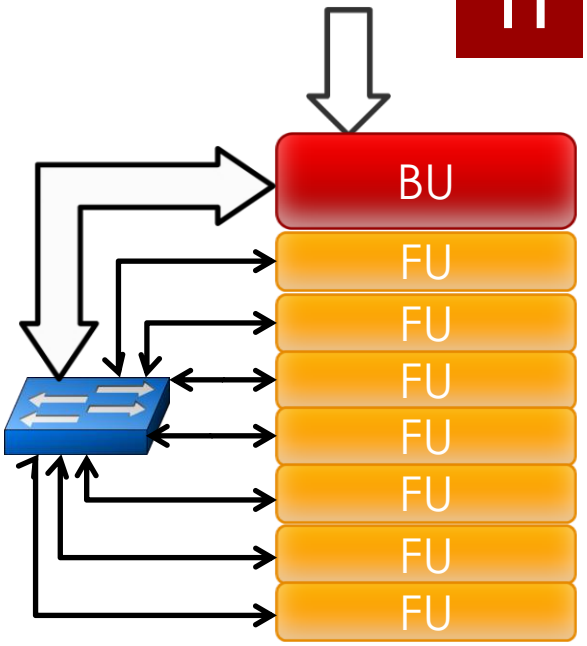
- Advantages:
  - Designed as a High Performance Computing interconnect over short distances (within datacenters)
  - Protocol is implemented in the network card silicon → low CPU load
  - 56 GBit/s per link (copper or optical)
  - Native support for Remote Direct Memory Access (RDMA)
  - No copying of bulk data between user space and kernel ('true zero-copy')
  - affordable
- Disadvantages:
  - Less widely known, API significantly differs from BSD sockets for TCP/IP
  - Fewer vendors than Ethernet
  - Niche market



Top500.org share by Interconnect family

File based filter farm and Storage

- In DAQ1, high level trigger process was running inside a DAQ application
  → introduces dependencies between online (DAQ) and offline (event selection) software which have different release cycles, compilers, state machines etc.

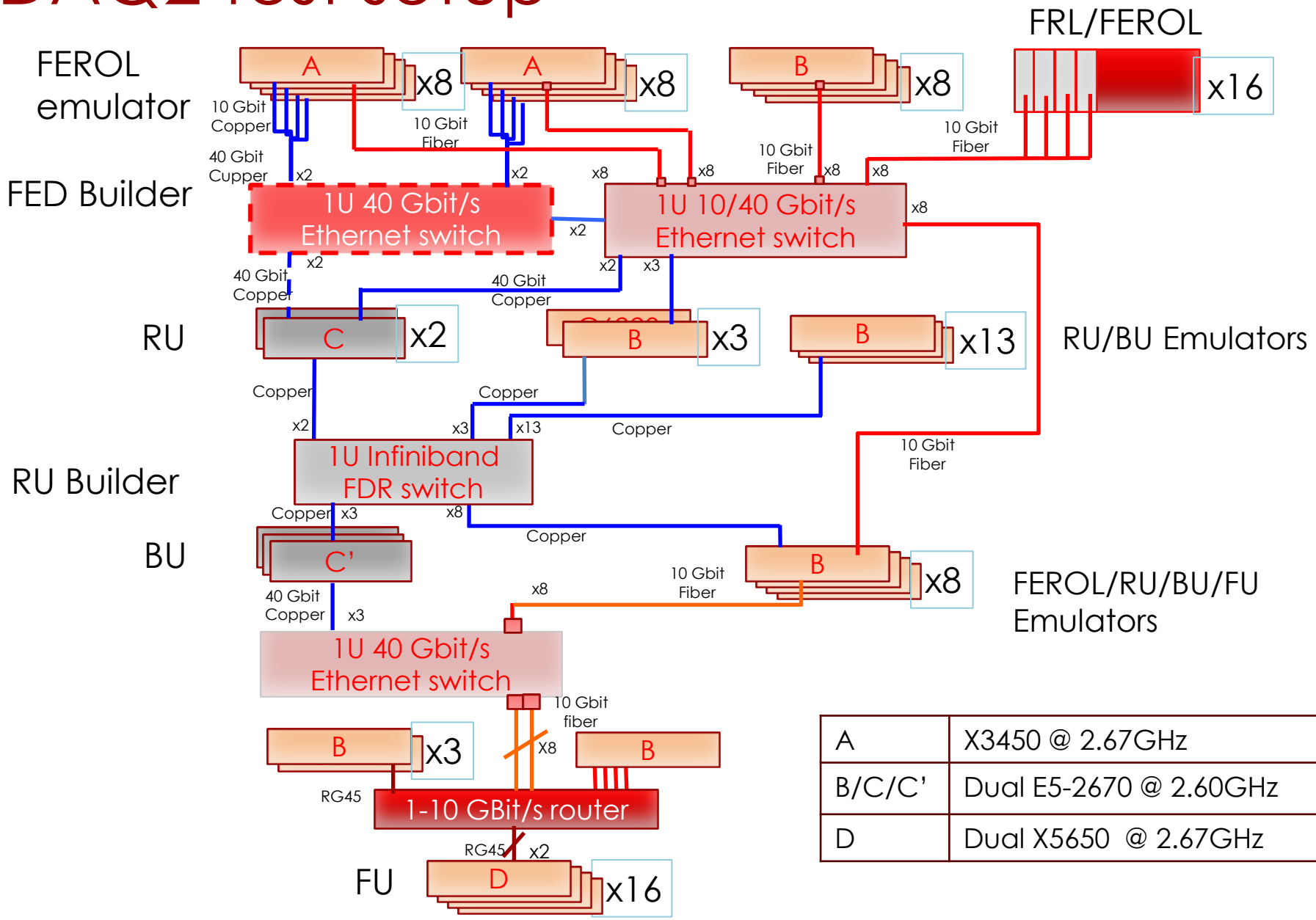


- Decoupling these needs a common, simple interface
  - files (no special common code required to write and read them)

see R. Mommsen's talk for more details

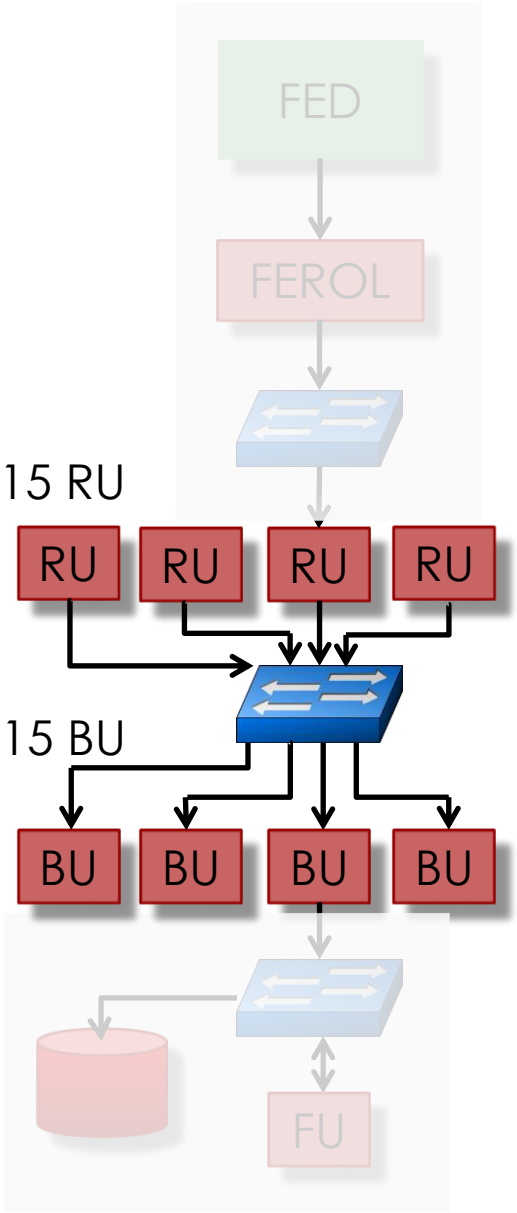
- Builder unit stores events in files in a RAM disk
  - Builder Unit acts as a NFS server, exports event files to Filter Unit PCs
  - Baseline: 2 Gbyte/s bandwidth
  - 'Local' within a rack
- Filter units write out selected events (~ 1 in 100) back to a global (CMS DAQ wide) filesystem (e.g. Lustre) for transfer to Tier0 computing centre

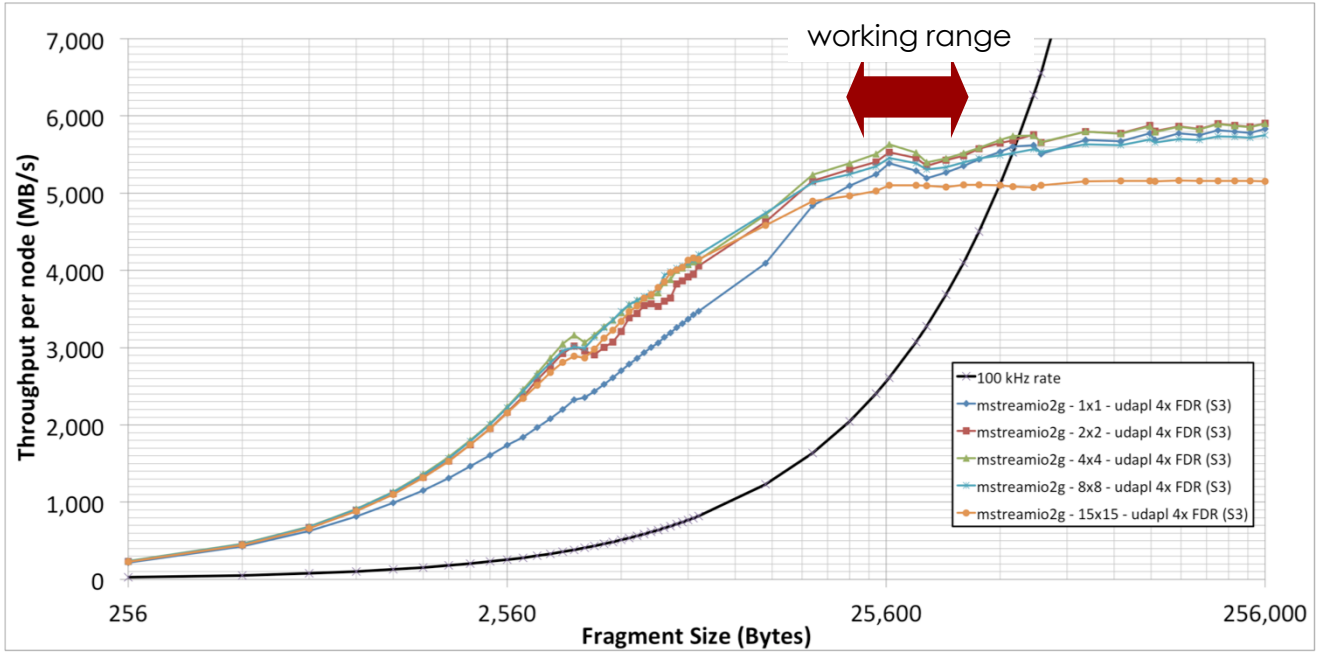
DAQ2 test setup



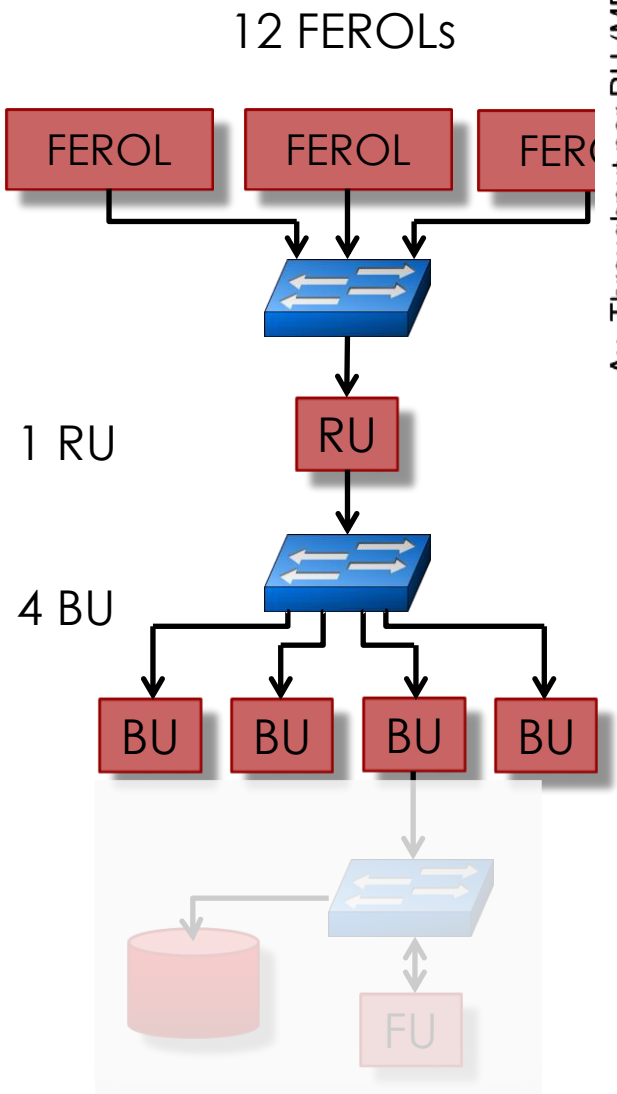
A	X3450 @ 2.67GHz
B/C/C'	Dual E5-2670 @ 2.60GHz
D	Dual X5650 @ 2.67GHz

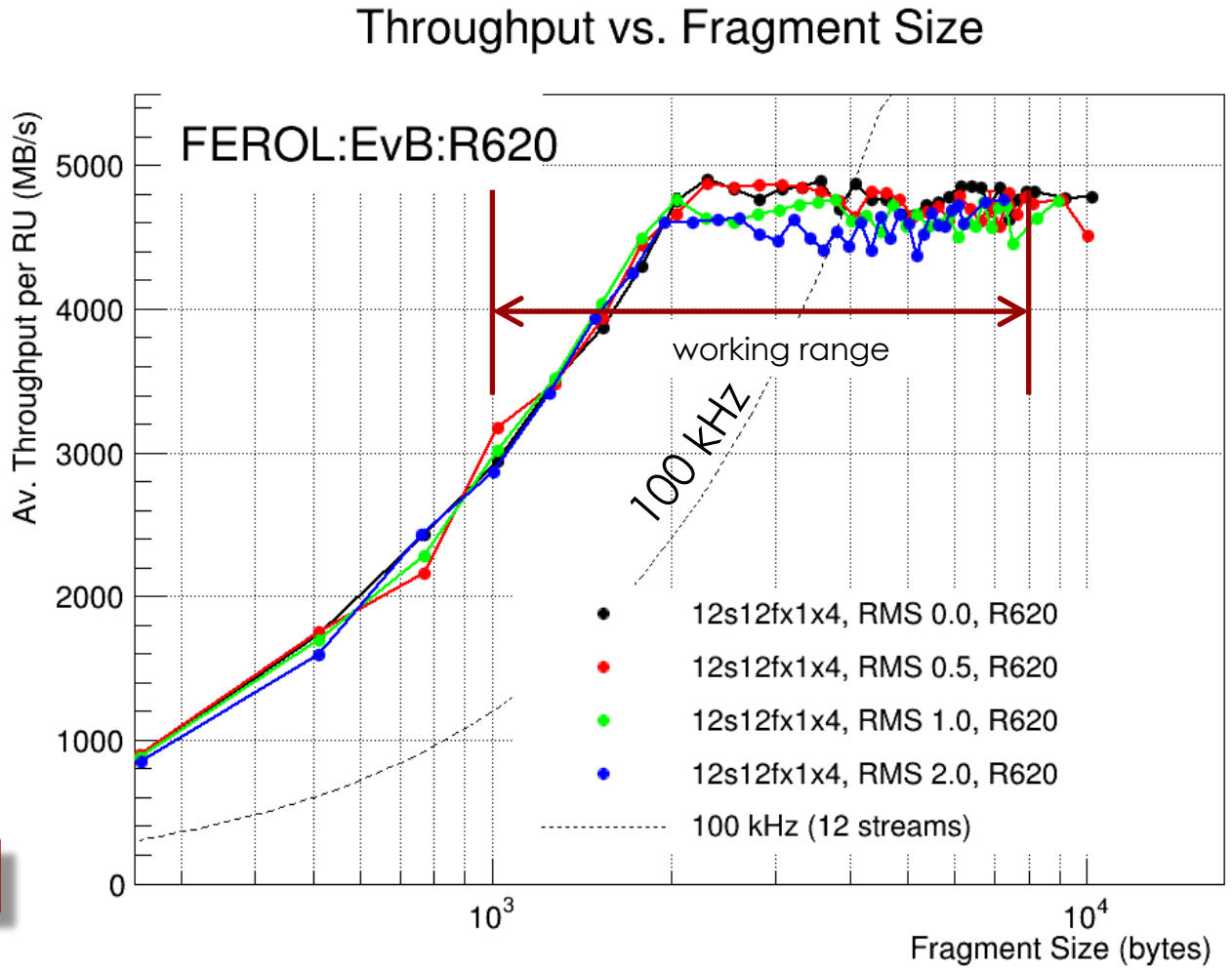
InfiniBand Measurements



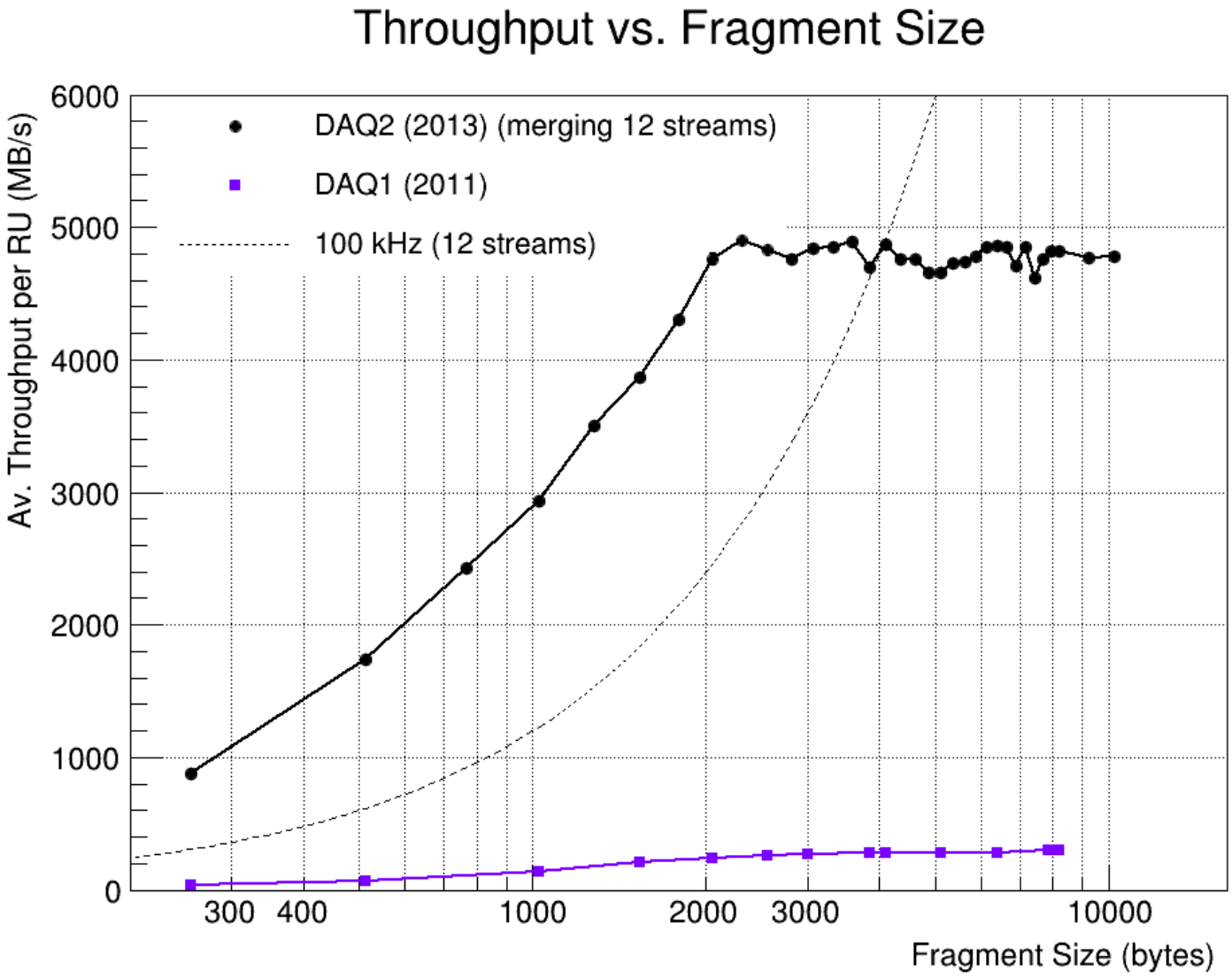


Test setup results





Test setup: DAQ1 vs. DAQ2



Comparison of throughput per Readout Unit

Summary / Outlook

- CMS has designed a central data acquisition system for post-LS1 data taking
  - replacing outdated standards by modern technology
  - ~ twice the event building capacity than DAQ system for Run1
  - accomodating a large dynamic range of up to 8 kByte fragments, flexible configuration
- Increase in networking bandwidth was faster than increase in event sizes
  - Number of event builder PCs reduced by a factor ~10
    - Each PC handles a factor ~10 more bandwidth
    - Requires performance related fine-tuning
- Performed various performance tests with a small scale demonstrator
- First installation activities for DAQ2 have started already
  - Full deployment foreseen for mid 2014
- Looking forward to recording physics data after the Long Shutdown 1 !